J Math Chem (2009) 46:1149–1157
DOI 10.1007/s10910-008-9511-3

ORIGINAL PAPER

A complexity-based measure and its application to phylogenetic analysis

Xiaoqi Zheng · Chun Li · Jun Wang

Received: 11 November 2007 / Accepted: 31 October 2008 / Published online: 9 December 2008

Abstract In this article, we propose two well-defined distance metrics of biological sequences based on a universal complexity profile. To illustrate our metrics, phylogenetic trees of 18 Eutherian mammals from comparison of their mtDNA sequences and 24 coronaviruses using the whole genomes are constructed. The resulting monophyletic clusters agree well with the established taxonomic groups.

Keywords Sequence complexity · mtDNA · SARS-CoV · Phylogenetic analysis

X. Zheng
Department of Applied Mathematics, Dalian University of Technology,
Dalian 116024,
People's Republic of China

X. Zheng
College of Advanced Science and Technology, Dalian University of Technology,
Dalian 116024,
People's Republic of China

C. Li
Department of Mathematics, Bohai University, Jinzhou 121000,
People's Republic of China

J. Wang
Department of Mathematics, Shanghai Normal University,
Shanghai 200234,
People's Republic of China

J. Wang (✉)
Scientific Computing Key Laboratory, Shanghai University,
Shanghai 200234,
People's Republic of China
e-mail: junwang@dlut.edu.cn

1 Introduction

The fast increase of many complete genomes of prokaryotes and eukaryotes raises a fundamental and challenging question to modern phylogenetics: how to reconstruct the phylogenetic history of different organisms using whole genomes? Traditional attempts require a multiple alignment of sequences and assume some sort of an evolutionary model. However, not to say the inherent computational complexity, it is meaningless to align two genomes because different genomes have different genes and gene order, and some evolutionary operations, such as rearrangements and lateral gene transfer, affect the final alignment seriously. Thus there is an urgent need to develop new sequence comparisons to deal with the ever increasing genome data.

Among the early attempts, Snel et al. [1] proposed *gene content* as a measure of similarity. The gene content between two sequences is defined as the number of genes they share divided by their total number of genes. This method is successful to compare long genomes for its light computational load, but fails to distinguish closely related species, e.g., mitochondrial (mt) genomes of placental mammals (all of them share the same genes and gene order). Observing that relative abundances of all dinucleotides are remarkably constant across the genome, Karlin et al. [2–4] proposed the "genome signature" to describe a genome. The genome signature consists of the array of dinucleotide relative abundances $\rho_{xy} = f_{xy}/f_x f_y$ extended over all dinucleotides, where f_x is the frequency of nucleotide x and f_{xy} is the frequency of dinucleotide xy. The final distance between two genomes is defined as the distance between their corresponding "signatures". Blaisdell [5] proposed a Markov chain model of biological sequences, and the difference between two sequences was quantified by the Euclidean distance between their transition matrices.

From the last decade of the 20th century, many data compression techniques, which were proved to be efficient in information storage and transmission, began to find their use in phylogenetic inferences [6–8]. The distance metric presented by Li et al. [6] is

$$d_K(S,T) = 1 - \frac{K(S) - K(S|T)}{K(ST)},$$

where $K(S|T)$ is the conditional Kolmogorov complexity of S given T, and $K(S)$ is the abbreviation of $K(S|\epsilon)$, with ϵ an empty string. However, Kolmogorov complexity is not a recursive function, that is, it is not incorporated in a computational scheme, and thus generally can only be approximated [9–11]. The complexity measure proposed by Lempel and Ziv [12–14] was an explicitly computable implementation of K-complexity for finite sequences, and many text compression algorithms were based on their measure (*gzip*, *zip*, and *Stacker*, for instance).

In the following text, we will first introduce the basic concepts and some properties regarding "LZ complexity", after which some previously proposed distance metrics are discussed. In the main text, two mathematically rigorous distance metrics based on "LZ-complexity" are proposed, and their applications are shown by constructing phylogenetic trees of 18 Eutherian mammals and 24 Coronaviruses including SARS-CoVs.

2 Methods and algorithms

2.1 Lempel-Ziv complexity

For symbol sequences S, T and R defined over a finite alphabet $\mathcal{A}$, let $l(S)$ be the length of S, $S(i)$ be the ith element of S and $S(i, j)$ be the subsequence of S that starts at position i and ends at position j. The sequence R is called an *extension* of S if R can be written as a concatenation of S and a given sequence T, i.e., $R = ST$.

An extension $R = ST$ from S is said to be *reproducible*, denoted by $S \rightarrow R$, if there exists an integer $p \leq l(S)$ such that $T(k) = R(p + k - 1)$, for $k = 1, \ldots, l(T)$. For example, WACC$\rightarrow$ WACCAC with $p = 2$, and AACGT$\rightarrow$ AACGTCGTC with $p = 3$. Moreover, if an extra different symbol at the end of the extension process is allowed, i.e., $S \rightarrow R(1, l(R) - 1)$, we can obtain the definition of *producible* extension, denoted by $S \Rightarrow R$. For example AACGT$\Rightarrow$AACGTCGTCW with $p = 3$. Thus we can say if $S \rightarrow R$ then $S \Rightarrow R$, but the reverse is not always true. An extension is called *exhaustive* if it is producible but not reproducible. For instance, the extension AACGT$\Rightarrow$AACGTCGTCW is exhaustive, but AACGT$\Rightarrow$AACGTCGTC is not.

According to the above definitions, any sequence S can be generated from the null sequence using iterated processes of "producible" extension. For example, the generating processes of S =AACGT can be written as: $\epsilon \Rightarrow$A$\Rightarrow$AA$\Rightarrow$AAC$\Rightarrow$AACG$\Rightarrow$ AACGT, or $\epsilon \Rightarrow$A$\Rightarrow$AAC$\Rightarrow$AACG$\Rightarrow$AACGT. The LZ complexity of a sequence S, denoted by $c(S)$, is defined as the minimum number of steps required to generate S from a null sequence using producible processes, e.g., c(AACGT)=4. It is easy to declare that, in this case, each extension is exhaustive with a possible exception of the last one, and the LZ complexity of any sequence is unique.

2.2 Distance metrics based on LZ complexity and their limitations

Note that $c(ST) - c(S)$ measures the amount of information in T when treating information in S as free. So if S and T are closely related to each other, $c(ST) - c(S)$ will be very small, i.e., it measures the degree of dissimilarity. From this consideration, Otu and Sayood [7] defined the distances between two sequences as follows:

$$d(S, T) = \max\{c(ST) - c(S), c(TS) - c(T)\}, \tag{1}$$

$$d^*(S, T) = c(ST) - c(S) + c(TS) - c(T). \tag{2}$$

(1), (2) and their normalized versions were used to infer the phylogenetic relationships among 20 Eutherian mammals. The resulting monophyletic clusters agree well with the established taxonomic groups. Li et al. [15] defined a *conditional producible operation*. The minimum number of conditional producible operations were considered as the conditional complexity given T, denoted by $c(S|T)$. The distance between two sequences S and T was characterized by

$$d_{\text{cond}}(S, T) = \max\{c(T|S), c(S|T)\}. \tag{3}$$

However, one can find that the above two distances are actually not metrics. Mathematically, a function $d(X, Y)$ defined on a set $\mathcal{S}$ is called a distance metric if for any $X, Y, Z \in \mathcal{S}$, the following three conditions are satisfied:

$$\text{Positivity \& Identity}: \ d(X, Y) \geq 0 \text{ and } d(X, Y) = 0 \Leftrightarrow X = Y; \tag{4}$$

$$\text{Symmetry}: d(X, Y) = d(Y, X); \tag{5}$$

$$\text{Triangle inequality}: \ d(X, Y) + d(Y, Z) \geq d(X, Z). \tag{6}$$

The above three conditions are essential characterizations of a distance measure. They are all necessary for the accurate clustering of a data set. But the above two distances do not satisfy the identity condition as $d(S, S) \neq 0$ if the last generating step is exhaustive, and $d(S, T)$ may be equal to 0 when $S \neq T$. Therefore, it is significant to revise the above distances or propose some more rigorous methods.

2.3 New distance metrics

In the present work, we define two new distance metrics of symbol sequences. Instead of considering the conditional compression ratio as done by Otu and Sayood, we make use of a maximum operation. The distances between two sequences S and T are defined as:

$$d_1(S, T) = \max_R\{|c(SR) - c(TR)|\}; \tag{7}$$

$$d_2(S, T) = \max_R\{|c(RS) - c(RT)|\}, \tag{8}$$

where R can be any sequence over the alphabet $\mathcal{A}$. The first distance is evaluated by the maximum complexity divergency between S and T when adding the same suffix. If change the suffix into prefix, we get $d_2(S, T)$.

We have two theorems below. Theorem 1 shows the validity of d_1 and d_2, and Theorem 2 gives a concrete computational approach of d_1.

Theorem 2.1 *The functions $d_1(S, T)$ and $d_2(S, T)$ are distance metrics, i.e., they satisfy the three axioms for a metric.*

Theorem 2.2 *Let S and T be any symbolic sequences. We have*

$$\begin{aligned} &\max\{c(ST) - c(TT), c(TS) - c(SS)\} \\ &\quad \leq d_1(S, T) \leq \max\{c(ST) - c(T), c(TS) - c(S)\}; \end{aligned} \tag{9}$$

$$\begin{aligned} &\max\{c(ST) - c(SS), c(TS) - c(TT)\} \\ &\quad \leq d_2(S, T) \leq \max\{c(ST) - c(S), c(TS) - c(T)\}. \end{aligned} \tag{10}$$

Proofs of above two theorems will be given in the Appendix. Note that $c(SS) = c(S)$ or $c(S) + 1$, for any sequence S. So the divergency between left and right hand side of $d_1(S, T)$ and $d_2(S, T)$ is 0 or 1. In the execution, we use their arithmetical average to approximate $d_1(S, T)$ and $d_2(S, T)$. Mind that the executive version

of d_2 is similar to the distance presented by Otu and Sayood (d). $d_2 = d$ when $c(ST) - c(S) > c(TS) - c(T)$ and $c(SS) = c(S)$, or $c(ST) - c(S) < c(TS) - c(T)$ and $c(TT) = c(T)$, while $d_2 = d - 0.5$ otherwise. So d_2 is actual a *refinement* of d. However, the difference between d and d_2 is so small, and perhaps no difference can be detected between their resulting phylogenies. In the following, we will use an alternative distance, d_1, to infer phylogenies of two data sets.

3 Applications to phylogenetic analysis

The mammalian phylogenetic relationship at the molecular level remains to be a controversial topic in nowadays molecular genetics. Different molecular data and analyses result in trees of different topological structures, and the most debatable is the relationship among three main groups of placental mammals, namely Primates, Ferungulates, and Rodents.

In the present work, we choose the whole mitochondrial genomes of 17 placental mammals and the platypus to construct the phylogenetic tree. Platypus, the only non-placental mammal, is selected as the outgroup. All the 18 data files are obtained from GenBank. In the first step, pairwise distance matrix is calculated using the distance d_1, then phylogenetic tree is constructed from the matrix using the UPGMA program in the PHYLIP package [16]. The final tree drawn by TREEVIEW [17] is shown in Fig. 1. According to our tree, species within each main group are clustered accordingly, and platypus stays outside of all 17 placental mammals. Notably, our result supports the topology of [Primates (Rodents, Ferungulates)], which is slightly different from the results of Li et al. [6] and Otu and Sayood [7].

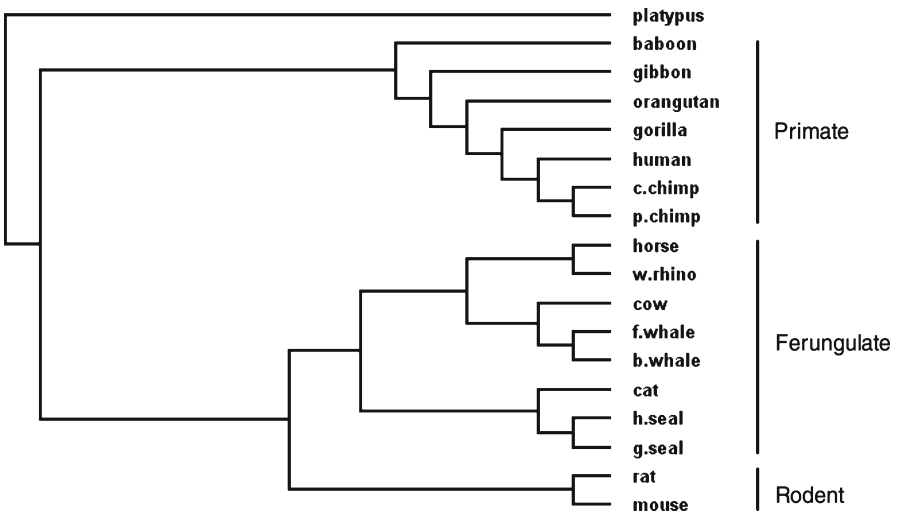

Fig. 1 Evolutionary tree of 18 mammalian species using the distance metric d_1. The resulting tree supports the topology of (Primates (Rodents, Ferungulates)), that is, Rodents and Ferungulates are the closest pair

Table 1 The accession number, abbreviation, name and length for each of the 24 coronavirus genomes

No.	Accession	Abbreviation	Genome	Group	Length(nt)
1	NC_002654	HCoV-229E	Human coronavirus 229E	I	27317
2	NC_002306	TGEV	Transmissible gastroenteritis virus	I	28586
3	NC_003436	PEDV	Porcine epidemic diarrhea virus	I	28033
4	U00735	BCoVM	Bovine coronavirus strain Mebuus	II	31032
5	AF391542	BCoVL	Bovine coronavirus isolate BCoV-LUN	II	31028
6	AF220295	BCoVQ	Bovin coronavirus strain Quebec	II	31100
7	NC_003045	BCoV	Bovine coronavirus	II	31028
8	AF208067	MHVM	Murine hepatitis virus strain ML-10	II	31233
9	AF201929	MHV2	Murine hepatitis virus stain 2	II	31276
10	AF208066	MHVP	Murine hepatitis virus stain Penn 97-1	II	31112
11	NC_001846	MHV	Murine hepatitis virus	II	31357
12	NC_001451	IBV	Avian infectiouis bronchitis virus	III	27608
13	AY278488	BJ01	SARS coronavirus BJ01	–	29725
14	AY278741	Urbani	SARS coronavirus Urbani	–	29727
15	AY278491	HKU-39849	SARS coronavirus HKU-39849	–	29742
16	AY278554	CUHK-W1	SARS coronavirus CUHK-W1	–	29736
17	AY282752	CUHK-Su10	SARS coronavirus CUHK-Su10	–	29736
18	AY283794	SIN2500	SARS coronavirus SIN2500	–	29711
19	AY283795	SIN2677	SARS coronavirus SIN2677	–	29705
20	AY283796	SIN2679	SARS coronavirus SIN2679	–	29711
21	AY283797	SIN2748	SARS coronavirus SIN2748	–	29706
22	AY283798	SIN2774	SARS coronavirus SIN2774	–	29711
23	AY291451	TW1	SARS coronavirus TW1	–	29729
24	NC_004718	TOR2	SARS coronavirus	–	29751

As another application, we infer the evolutionary relationships of 24 coronaviruses including SARS-CoVs (Severe Acute Respiratory Syndrome). Coronaviruses are members of a family of enveloped viruses that replicate in the cytoplasm of animal host cell. According to the type of the host, coronaviruses isolated previously are classified into three groups, groups I and II contain mammalian viruses, whereas group III contains only avian viruses. Marra et al. [18] and Rota et al. [19] first chose data from above three groups and some SARS-CoVs to construct phylogenetic tree. Their results indicate that SARS-CoVs are not closely related to any of the previously characterized coronaviruses and form a new fourth group. Using similar data, Zheng et al. [20] applied a geometric approach. They transformed each coronavirus genome into "Z-curve", an equivalent graphical representation of DNA sequence, then used geometric center and three eigenvectors of "Z-curve" as descriptors of this genome.

Motivated by Rota et al., Marra et al. and Zheng et al., we use our distances to infer the phylogenetic relations of the above coronaviruses (data are shown in Table 1). However, different from the first data set (mitochondrial genomes), lengths of coronavirus genomes vary significantly (from 27 kb to above 31 kb). In order to eliminate the effects of different lengths, we normalize our distances by dividing the sum and maximum of $c(S)$ and $c(T)$. The consensus tree drawn by TREEVIEW is shown as Fig. 2. We find that our topology coincides well with the conventional taxonomic groups, i.e., coronaviruses within each typical group (I human coronaviruses, II bovine coronaviruses and Murine hepatitis viruses, III avian viruses) cluster accordingly, and

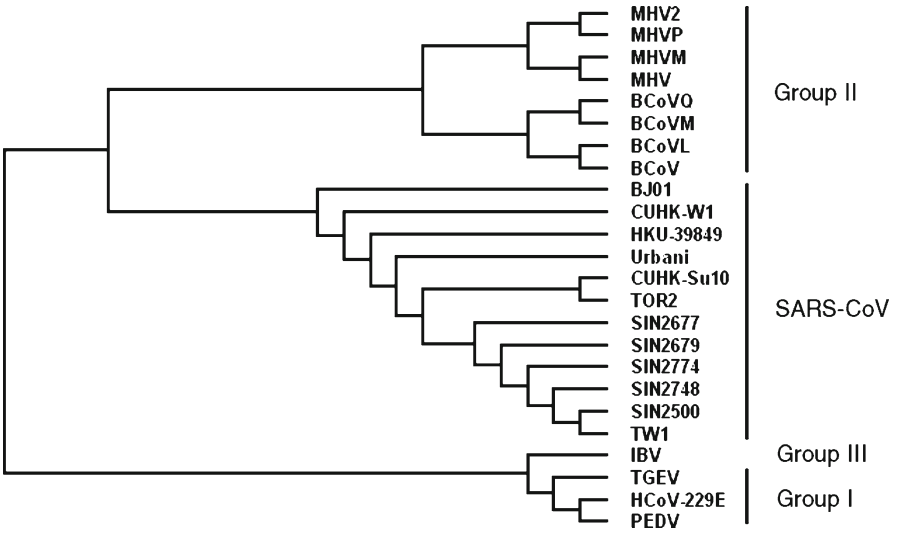

Fig. 2 The consensus tree constructed by two normalized versions of the distance d_1 (by dividing the maximum and sum of $c(S)$ and $c(T)$, respectively). According to this tree, all 12 SARS-CoVs are clustered and form new group, which is distantly related to the group II coronaviruses

all 12 SARS-CoV strains are grouped together and form a new fourth group, which is distantly related to the group II coronaviruses. This result is in accordance with maximum likelihood tree built from a fragment of the spike protein [21], and also agrees with the alignment tree from comparing replicase ORF1b amino acid sequences of some viruses [22].

4 Conclusions

With the completion of many genome projects of Prokaryotes and Eukaryotes, whole genome phylogenies are available and expected to be more accurate compared to traditional experiments on only a single gene or a fragment of genome. However, multiple sequence alignment of genomic sequences is still a bottleneck, first due to the computational time, and second due to the inherent model assumptions. Therefore, there is a great need to develop new sequence comparisons free of the above problems. In recent years, a quantity of alignment-free methods, e.g. complexity-based approaches [6,7], k-words composition [23] and graphical representations [24–30] have been proposed. However, compared to the alignment method, alignment-free comparison methods are still in their premature stage.

In this article, we propose two well-defined distance metrics on the basis of a universal sequence complexity. To illustrate them, phylogenetic trees of 18 mammals and 24 coronaviruses are constructed, and results show that our distances can successfully cluster species at different levels. In the first data set, our tree supports the topology of [Primates (Rodents, Ferungulates)], i.e., Rodents and Ferungulates are the closest pair. Phylogenetic tree built from the second data set shows that all 12

SARS-CoV strains are grouped together and form a new fourth group, which is distantly related to the group II coronaviruses. In contrast to the traditional alignment method, our method does not suffer from some evolutionary operations, e.g., gaps and large rearrangements in genomic sequences. Moreover, it is fully automated, i.e., does not need any free parameter and human intervention. So it could serve as an alternative way of genome comparison, especially in the case that there are no agreed-upon evolutionary models.

Acknowledgements This work was supported in part by Leading Academic Discipline Project of Shanghai Normal University (No. DZL803) and Shanghai Leading Academic Discipline Project (No. S30405).

Appendix

Proof of Theorem 1 The positivity and symmetry conditions obviously hold according to the definition of d_1. We will check the triangle condition by introducing a sequence Q.

$$\begin{aligned} d_1(S,T) &= \max_R\{|c(SR)-c(TR)|\} \\ &= \max_R\{|c(SR)-c(QR)+c(QR)-c(TR)|\} \\ &\le \max_R\{|c(SR)-c(QR)|+|c(QR)-c(TR)|\} \\ &\le \max_{R_1}\{|c(SR_1)-c(QR_1)|\}+\max_{R_2}\{|c(QR_2)-c(TR_2)|\} \\ &= d_1(S,Q)+d_1(Q,T). \end{aligned}$$

which is the triangle inequality, as required. We now need to show that $d_1(S,T)$ satisfies the identity condition. By definition, if $S=T$, then $d_1(S,T)=d_1(S,S)=\max_R\{|c(SR)-c(SR)|\}=0$. In the following, we will prove that $d_1(S,T)\neq 0$ if $S\neq T$. This assertion is clear when $c(S)\neq c(T)$ (we only set R be null sequence). When $c(S)=c(T)$, we have three cases to be considered:

(1) If one of S's and T's last generating steps is exhaustive, we can add a letter l to both ends of S and T, so that $c(Sl)=c(Tl)\pm 1$, yielding $d_1(S,T)\ge 1$.
(2) The last generating steps of S and T are both exhaustive. Without loss of generality, we select a subsequence R of S, which does not appear in sequence T. Then we have $c(SRl)=c(S)+1$, and $c(TRl)>c(T)+1$, where l is an arbitrary letter from alphabet $\mathcal{A}$. So $d_1(S,T)\ge|c(SRl)-c(TRl)|\ge 1$.
(3) If neither of the last generating step of S and T is exhaustive, we can add the same suffix Q to these two sequences till at least one of the last steps of SQ and TQ is exhaustive, which will come back to the above two cases.

In order to prove Theorem 2, we first give a lemma:

Lemma 1 *$c(STR)-c(ST)\le c(TR)-c(T)$, for any sequences S, T and R.*

Proof The left hand side is the number of components R would have when parsed using ST, and the right hand side is the number of components R would have when

parsed using S. Having ST instead of T cannot increase the number of components in parsing of R. So the inequality holds. □

Proof of Theorem 2 Put R be T in Formula (7). Then $c(ST)-c(TT) \leq \max_R\{c(SR)-c(TR)\}$. Similarly, $c(TS)-c(SS) \leq \max_R\{c(TR)-c(SR)\}$. So $\max\{(ST)-c(TT), c(TS)-c(SS)\} \leq \max_R\{|c(SR)-c(TR)|\}$, which is the left half of Inequality (9). We now prove the right half of Inequality (9). According to Lemma 1, $c(STR)-c(TR) \leq c(ST)-c(T)$. Note that $c(STR) \geq c(SR)$, so $c(SR)-c(TR) \leq c(ST)-c(T)$, for any sequences S, T and R. Similarly, $c(TR)-c(SR) \leq c(TS)-c(S)$. In conclusion, $\max\{c(SR)-c(TR), c(TR)-c(SR)\} \leq \max\{c(ST)-c(T), c(TS)-c(S)\}$, i.e., $d_1(S,T) \leq \max\{c(ST)-c(T), c(TS)-c(S)\}$.

The proofs for d_2 are analogous to the case of d_1.

References

1. B. Snel, P. Bork, M.A. Huynen, Nat. Genet. **21**, 108 (1999)
2. A. Campbell, J. Mrazek, S. Karlin, Proc. Natl. Acad. Sci. USA **96**, 9184 (1999)
3. S. Karlin, I. Ladunga, Proc. Natl. Acad. Sci. USA **91**, 12832 (1994)
4. S. Karlin, J. Mrázek, Proc. Natl. Acad. Sci. USA **94**, 10227 (1997)
5. B.E. Blaisdell, Proc. Natl. Acad. Sci. USA **83**, 5155 (1986)
6. M. Li, J.H. Badger, X. Chen, S. Kwong, P. Kearney, H.Y. Zhang, Bioinformatics **17**, 149 (2001)
7. H.H. Otu, K. Sayood, Bioinformatics **19**, 2122 (2003)
8. C. Li, J. Wang, Similarity analysis of DNA sequences based on the generalized LZ complexity of (0,1)-sequences, Preprint, J. Math. Chem. (2006)
9. T.M. Cover, J.A. Thomas, *Elements of Information Theory* (Tsinghua University Press: Beijing, 2003)
10. Y.L. Orlov, V.N. Potapov, Nucleic Acids Res. **32**, 628 (2004)
11. T. Jiang, Y. Xu, M.Q. Zhang, *Current Topics in Computational Molecular Biology* (Tsinghua University Press and The MIT Press, Tsinghua and Cambridge, 2002), pp. 157–171
12. A. Lempel, J. Ziv, IEEE T. Inform. Theory **22**, 75 (1976)
13. J. Ziv, A. Lempel, IEEE T. Inform. Theory **23**, 337 (1977)
14. J. Ziv, A. Lempel, IEEE T. Inform. Theory **24**, 530 (1978)
15. B. Li, Y.B. Li, H.B. He, Geno. Prot. Bioinfo. **3**, 206 (2005)
16. J. Felsenstein, PHYLIP (Phylogeny Inference Package) version 3.5c. Department of Genetics, University of Washington, Seattle (1993)
17. R.D. Page, Comput. Appl. Biosci. **12**, 357 (1996)
18. M.A. Marra et al., Science **300**, 1399 (2003)
19. P.A. Rota et al., Science **300**, 1394 (2003)
20. W.C. Zheng, L.L. Chen, H.Y. Ou, F. Gao, C.T. Zhang, Mol. Phylogenet. Evol. **36**, 224 (2005)
21. P. Liò, N. Goldman, Trends Microbiol. **12**, 106 (2004)
22. E.J. Snijder, P.J. Bredenbeek, J.C. Dobbe, V. Thiel, J. Ziebuhr, L.L. Poon, Y. Guan, M. Rozanov, W.J. Spaan, A.E. Gorbalenya, J. Mol. Biol. **331**, 991 (2003)
23. S. Vinga, J. Almeida, Bioinformatics **19**, 513 (2003)
24. A. Nandy, Curr. Sci. **66**, 309 (1994)
25. A. Nandy, P. Nandy, Chem. Phys. Lett. **368**, 102 (2003)
26. M. Randić, A.T. Balaban, M. Novič, A. Založnik, T. Pisanski, Period. Biol. **107**, 403 (2005)
27. M. Randić, D. Butina, J. Zupan, Chem. Phys. Lett. **419**, 528 (2006)
28. R. Zhang, C.T. Zhang, J. Biomol. Struc. Dyn. **11**, 767 (1994)
29. B. Liao, R. Li, W. Zhu, X. Xiang, J. Math. Chem. **42**, 47 (2007)
30. Y.S. Zhang, M.S. Tan, Visualization of DNA sequences based on 3DD-Curves, Preprint, J. Math. Chem. (2007)